Application of an Ensemble-Based Medical Visual Question Answering (Med-VQA) System for Analyzing Polyps in Gastrointestinal Endoscopy Images

Wai-Shing Ng
Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
wsngam@connect.ust.hk

Abstract

Gastrointestinal cancer represents the third most common cancer globally, with significant morbidity and mortality linked to polyps that can transition from benign to malignant. Current screening methods, primarily gastrointestinal endoscopy, involve substantial manual analysis of images, emphasizing the need for automated systems to enhance efficiency and accuracy in polyp classification. This paper presents an Ensemble-Based Medical Visual Question Answering (Med-VQA) system designed to integrate computer vision and natural language processing for improved interaction between healthcare professionals and diagnostic tools. By leveraging an ensemble of diverse spatial models—specifically ResNet50 and VGG16—and feature fusion techniques such as concatenation and element-wise multiplication, we optimize the diagnostic output from endoscopic images in response to clinician queries. Our experiments demonstrate that the Med-VQA system achieves an accuracy of 83.64% *with the ResNet50 architecture and element-wise multiplication method, while ensemble inference yields enhanced performance, reaching an accuracy of* 85.92%. *This study underscores the potential of integrating spatial and linguistic information to foster deeper insights and interactivity in the clinical setting, consequently supporting early detection and prevention strategies for gastrointestinal cancer.*

1. Introduction

According to the World Health Organization (WHO), gastrointestinal cancer is the third most prevalent cancer globally, with over 1.9 million new cases of gastrointestinal cancer and more than 930,000 associated deaths reported in 2020 [12]. This type of cancer often originates from polyps, which are small cell clusters that develop within the gastrointestinal tract [10]. Although the majority of polyps are benign, a significant proportion (between 5% and 10%) can progress to cancer over time [10]. Due to the asymptomatic nature of polyps, regular health screening is crucial for early detection and prevention of gastrointestinal cancer.

Gastrointestinal endoscopy is a widely utilized screening technique involving the insertion of a flexible endoscope through a patient's mouth or anus to provide real-time visualization of the gastrointestinal tract's mucosal lining and surrounding tissues. However, analyzing the resulting images is labor-intensive, requiring healthcare professionals to manually examine large volumes of endoscopic images and classify whether the polyps are cancerous [9]. In addition, polyps exhibit significant variability in size, shape, and color, causing smaller or flatter polyps particularly challenging to identify compared to larger ones [9]. These challenges highlights the necessity for an automated polyp classification system to enhance screening efficiency and accuracy.

With advancements in artificial intelligence (AI), researchers have developed machine learning models, such as Convolutional Neural Networks (CNN) [1] and VGG [3], for the automatic classification of polyps in gastrointestinal images. Despite their proficiency in polyp classification [1, 3], these models often lack interactive capabilities with healthcare professionals, limiting their utility in clinical settings. To address this limitation, Med-VQA systems present a promising solution. These systems integrate computer vision (CV) to interpret endoscopic images and employ natural language processing (NLP) to analyze queries from healthcare professionals [7].

In this paper, we propose a development on an Ensemble-Based Med-VQA System to analyze polyps in gastrointestinal endoscopy images, leveraging spatial and linguistic feature integration for generating accurate diagnostic outputs. When healthcare professionals pose questions, the Ensemble-Based Med-VQA system generates answers based on the combination of the image features and the query context. This approach not only enhances the depth of insights derived from the images but also improves the interactivity of the experience for healthcare professionals. In summary, our main contributions are:

- Conducting experiments with diverse spatial models (ResNet50, VGG16) and feature fusion algorithm (Concatenation, Element-wise Multiplication) to identify the optimal architecture for the Med-VQA system.
- Evaluating the effect of various weight decay values for regularization on model performance to optimize the Med-VQA system's efficacy.
- Integrating the top-performing Med-VQA systems to for ensemble inference.

Our results indicate that the Med-VQA system, implemented with the ResNet50 architecture and the element-wise multiplication algorithm in conjunction with weight decay regularization, achieves superior performance, with an accuracy of 83.64%, compared to other model architectures. Furthermore, the application of the ensemble inference method further enhances the model's performance, resulting in an accuracy of 85.92%.

2. Related Work

Polyp Detection. In the study by Dmitrii et al. [10], Convolutional Neural Network (CNN) architectures were utilized to successfully detect polyps in gastrointestinal endoscopy images. By leveraging the hierarchical feature extraction capabilities of CNN, their approach demonstrated proficiency of CNN architecture in capturing critical features of polyps and significantly improving detection accuracy. Notwithstanding these advancements, Ali et al. [12] pointed out that current systems primarily concentrate on detecting the presence of polyps but often struggle to accurately predict their malignancy, location, and size. This information is vital for healthcare professionals to make informed decision regarding patient management, including treatment strategies and cancer evaluation.

To address these challenges, we propose the development of a Med-VQA system designed to provide healthcare workers with comprehensive diagnostic information through an interactive approach, thereby enhancing decision-making in patient care. Furthermore, building upon the advancements in CNN architectures demonstrated by Dmitrii et al. [10], we will leverage models such as ResNet50 and VGG16 to effectively capture the features of polyps within our VQA system.

Med-VQA System. Thai et al. [11] established essential methodologies for Med-VQA, employing CNNs for image feature extraction and the BERT model for producing contextualized word embeddings that capture nuanced linguistic features critical for understanding complex medical queries. A concatenation algorithm is then utilized to combine these features, with the integrated output processed through a classification model. However, this standard feature fusion algorithm hinders the interpretability of feature representations and increases dimensionality, complicating the learning process and potentially degrading model performance.

To address these challenges, we propose the implementation of an element-wise multiplication (EM) algorithm within our Med-VQA system. This approach facilitates more effective integration of image and text features, enhancing interaction between visual and textual modalities while preserving the dimensionality of the combined features.

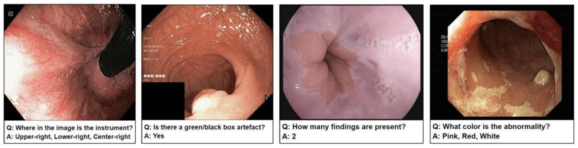
Ensemble Learning. While recent studies [10–12] have employed single-model systems, Bertels et al. [2] and Li et al. [6] have identified inherent biases and sensitivity to outliers in these approaches. These challenges have prompted the exploration of ensemble methodologies that integrate multiple learning algorithms for more robust predictions. Ensemble learning capitalizes on the strengths of individual models to mitigate the limitations of single-model systems [2]. By aggregating predictions from various algorithms trained on identical datasets or different subsets, these methods significantly reduce variance, enhance generalization, and improve overall performance [6].

Given these insights, integrating ensemble learning strategies into our Med-VQA configurations has the potential to enhance the reliability and accuracy of polyp detection in gastrointestinal endoscopic images.

3. Dataset

3.1. Description

ID	Questions	Sample Answers
0	What type of procedure is the image taken from?	"Colonoscopy", "Gastroscopy"
1	Have all polyps been removed?	"Yes", "No", "Not relevant"
2	Is this finding easy to detect?	"Yes", "No", "Not relevant"
3	Is there a green/black box artifact?	"Yes", "No"
4	Is there text?	"Yes", "No"
5	What color is the abnormality?	"Red", "Pink", "White, Yellow", ...
6	What color is the anatomical landmark?	"Red", "Red, White", "Pink, Red, grey", ...
7	How many findings are present?	"0", "1", "2", "3", "4", "5", ...
8	How many polyps are in the image?	"0", "1", "2", "3", "4", "5", ...
9	How many instruments are in the image?	"0", "1", "2", "3"
10	Where in the image is the abnormality?	"Center", "Lower-left", "Lower-right, Center-right", ...
11	Where in the image is the instrument?	"Center", "Lower-left", "Lower-right, Center-right", ...
12	Are there any abnormalities in the image?	"No", "Polyp", "Ulcerative colitis", "Oesophagitis", ...
13	Are there any anatomical landmarks in the image?	"No", "Z-line" "Cecum", "Ileum", "Pylorus", "Not relevant"
14	Are there any instruments in the image?	"No", "Tube", "Biopsy forceps", "Metal clip", 'Polyp snare, Tube', ...
15	Where in the image is the anatomical landmark?	"Center", "Lower-left", "Lower-right, Center-right", ...
16	What is the size of the polyp?	"< 5mm", "5-10mm", "11-20mm", ">20mm", "Not relevant", ...
17	What type of polyp is present?	"Paris ip", "Paris iia", "Paris is", "Paris is, Paris iia", ...


Q: Where in the image is the instrument?
A: Upper-right, Lower-right, Center-right
Q: Is there a green/black box artefact?
A: Yes
Q: How many findings are present?
A: 2
Q: What color is the abnormality?
A: Pink, Red, White

(a) No abnormality (b) Polyp (c) Oesophagitis (d) Ulcerative Colitis

Figure 1. Illustrations of question-answer pairs along with common abnormalities in gastrointestinal images from the ImageCLEFmed-Med-VQA-GI-2023 dataset. Source: [11]

We utilized the Med-VQA-GI dataset [5] from the ImageCLEF 2023 challenge, which comprises 2,000 RGB gastrointestinal endoscopy images with varying dimensions. Each image is linked to 18 textual questions related to abnormalities, surgical instruments, normal findings, and other artifacts, with the potential for multiple answers per question, as shown in Figure 1. The 36,000 textual question-answer pairs are structured within the json file using the imageID as the key. We will employ an 80-10-10 train-validation-test split, allocating 1,600 images for model training, 200 images for validation, and 200 images for testing.

3.2. Data Preprocessing

3.2.1 Question-Answer Pairing

As discussed in Section Sec. 3.1, each image is associated with 18 questions, each of which may contain multiple answers. To streamline model training and evaluation, we associate each image with a single question-answer pair, repeated 18 times to address all relevant questions. Furthermore, answers comprising multiple responses to a question will be encoded as a single categorical variable to simplify the classification process. For example, as shown in Figure 1, if the question ”What color is the abnormality?” has possible answers ”Pink,” ”Red,” and ”White,” the answers will be encoded as a single category: ”Pink, Red, White.”

3.2.2 Image Resizing

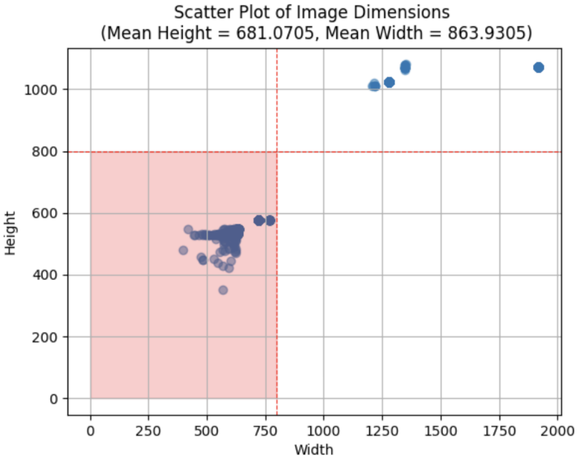


Figure 2. Scatter Plot of Image Dimensions. The red region highlights that most of the image has both width and height less than 800 pixels. Compromising between the dataset size and the image resolution, we chose to resize the image into a consistent dimension of 800x800 pixels

To address the varying dimensions in the dataset, as detailed in Sec. 3.1, we employed image resizing to ensure compatibility with our model. An analysis of the training image dimensions (see Figure 2) indicates that most images have both width and height measurements below 800 pixels. To balance dataset size and image resolution, we adopted a uniform target dimension of 800x800 pixels for all images. Prior to model input, all images will be resized to this standardized dimension using linear interpolation, which promotes accurate alignment and effective processing within the model.

3.2.3 Image Normalization

Image normalization is a critical preprocessing step applied to resized images to enhance the stability and efficiency of the training process. This method standardizes the pixel values across each channel, targeting a mean of 0 and a standard deviation of 1, which facilitates stable and accelerated convergence during model training. We begin by analyzing the distribution of pixel values within the training dataset, followed by normalization using the equation:

$$x_{\text{normalized},c} = \frac{x_c - \mu_c}{\sigma_c} \quad \text{for } c \in \{R, G, B\} \tag{1}$$

where:

- $\mu_R = 141.45$, $\mu_G = 93.52$, $\mu_B = 85.09$
- $\sigma_R = 80.83$, $\sigma_G = 61.44$, $\sigma_B = 61.27$

4. Methods

4.1. Image Feature Extraction

Each image feature $x_{\text{Image}} \in \mathbb{R}^{800\times800\times3}$ is transformed into an image embedding $\hat{x}_{\text{image}} \in \mathbb{R}^{\text{ImageEmbeddingDim}}$ using an image encoder, as illustrated in Figure 3. This transformation is denoted by $G(\cdot)$, allowing us to express the image embedding as follows:

$$\hat{x}_{\text{image}} = G(x_{\text{image}}) \tag{2}$$

We have developed various Med-VQA systems utilizing the following two spatial models.

4.1.1 VGG16

VGG16 [8] is a CNN architecture recognized for its proficiency in image feature extraction. Its design consists of sequential convolutional layers followed by max pooling layers, which facilitate hierarchical feature extraction and effectively capture both low-level and high-level features. The default output dimension for VGG16 in the torchvision library is configured to (BatchSize, 512, 7, 7). To ensure seamless integration of image and text features at the fusion layer, an Adaptive Average Pooling layer (AdaptiveAvgPool2d) is applied, restructuring the output dimensions to

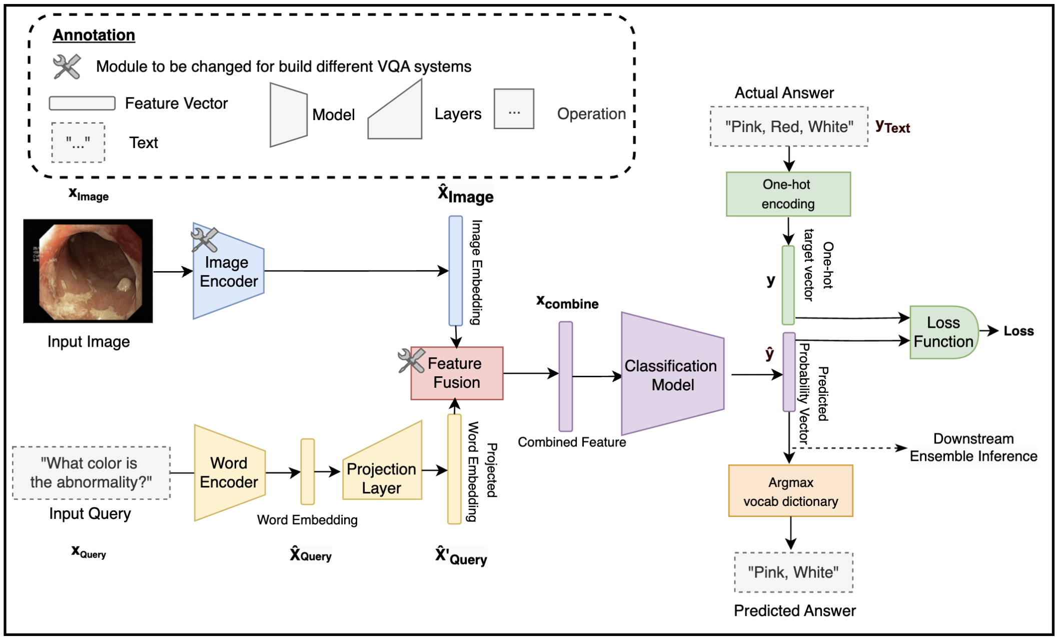


Figure 3. Overview of the Med-VQA system, serving as a template for developing various Med-VQA configurations with different combinations of image encoders and feature fusion algorithms. The system processes an RGB input image alongside a textual input query to produce a predicted textual answer. Additionally, it computes the loss for backpropagation and generates a predicted probability vector for downstream ensemble inference.

(BatchSize, 512). Consequently, the resulting tensor is represented as $\hat{x}_{\text{image}} \in \mathbb{R}^{512}$.

4.1.2 ResNet50

ResNet-50 [4] is another deep CNN architecture renowned for its efficacy in image feature extraction. Despite comprising 50 layers, this model employs residual connections within each block to enhance gradient flow by allowing the gradient to bypass one or more layers. This design enables the training of deeper networks without performance degradation. The default output dimension for ResNet-50 in the torchvision library is set to (BatchSize, 2048), facilitating seamless integration with one-dimensional word embeddings within the fusion layer. Consequently, we represent the image embedding as $\hat{x}_{\text{image}} \in \mathbb{R}^{2048}$.

4.2. Textual Feature Extraction

Each textual input x_{Query} is tokenized and transformed into a word embedding $\hat{x}_{\text{query}} \in \mathbb{R}^{\text{WordEmbeddingDim}}$ using a word encoder, as depicted in Figure 3. Among the Med-VQA systems we have developed, the BERT [8] model is employed as the word encoder for the word-to-feature transformation, leveraging its advanced contextual understanding and ability to capture nuanced linguistic features. This ensures consistent tokenization and embedding of words. Denote this transformation by $H(\cdot)$, so the word embedding is represented as:

$$\hat{x}_{\text{query}} = H(x_{\text{query}}) \tag{3}$$

The BERT model is configured with a fixed TextEmbeddingDim = 768 in our Med-VQA model setup. To facilitate subsequent feature processing, a projection layer is employed to map the textual word embedding to the dimension of the image embedding, s.t. $\hat{x'}_{\text{query}} \in \mathbb{R}^{\text{ImageEmbeddingDim}}$. Let this transformation be denoted by Proj, resulting in a projected word embedding:

$$\hat{x'}_{\text{query}} = \text{Proj}(\hat{x}_{\text{query}}) \tag{4}$$

4.3. Feature Fusion

After that, the image embedding $\hat{x}_{\text{image}} \in \mathbb{R}^{\text{ImageEmbeddingDim}}$ and the projected word $\hat{x'}_{\text{word}} \in \mathbb{R}^{\text{ImageEmbeddingDim}}$ embedding will be fused together into a combined embedding $x_{\text{combine}} \in \mathbb{R}^{\text{CombineEmbeddingDim}}$ for the later classification to generate the predicted answer as shown in Figure 3 . We have build various models using the following two algorithms.

4.3.1 Concatenation

This is a simple feature fusion algorithm that appends the projected textual feature $\hat{x'}_{\text{word}} \in \mathbb{R}^{\text{ImageEmbeddingDim}}$ to the image feature $\hat{x}_{\text{image}} \in \mathbb{R}^{\text{ImageEmbeddingDim}}$ into a combined feature $x_{\text{combine}} \in \mathbb{R}^{2\times\text{ImageEmbeddingDim}}$, such that:

$$x_{\text{combine}} = \left[\hat{x}_{\text{image}}, \hat{x'}_{\text{word}}\right] \tag{5}$$

4.3.2 Element-wise Multiplication

In our approach, element-wise multiplication is utilized to integrate the image feature $\hat{x}_{\text{image}} \in \mathbb{R}^{\text{ImageEmbeddingDim}}$ with the projected textual feature $\hat{x'}_{\text{word}} \in \mathbb{R}^{\text{ImageEmbeddingDim}}$. This operation produces a combined feature vector $x_{\text{combine}} \in \mathbb{R}^{\text{ImageEmbeddingDim}}$. Element-wise multiplication, denoted by the Hadamard product symbol $\odot$, is defined as follows:

$$x_{\text{combine}} = \hat{x}_{\text{image}} \odot \hat{x'}_{\text{word}} \tag{6}$$

This operation is equivalent to computing each component i as:

$$x^{(i)}_{\text{combine}} = \hat{x}^{(i)}_{\text{image}} \cdot \hat{x'}^{(i)}_{\text{word}} \tag{7}$$

where $x^{(i)}$ denotes the i-th element of the vector.

4.4. Classification

The combined feature $x_{\text{combine}} \in \mathbb{R}^{\text{CombineEmbeddingDim}}$ is processed through the classification model to generate the predicted probability vector, thereby producing answers for the user. Our Med-VQA system utilizes a classification model based on a multilayer perceptron (MLP) as shown in Figure 3.

Layer	Output Shape
Input Layer (units = C.E.Dim)	(C.E.Dim,)
Dropout Layer (rate = 0.3)	(C.E.Dim,)
Dense Layer (units = 2048)	(2048,)
LeakyReLU (slope = 0.01)	(2048,)
Dense Layer (units = 1024)	(1024,)
LeakyReLU (slope = 0.01)	(1024,)
Output Layer (units = #AnsToken)	(#AnsToken,)
Softmax	(#AnsToken,)

Table 1. Overview of the Multilayer Perceptron (MLP) architecture used for classification in the Med-VQA system.

As shown in Table 1, our Multilayer Perceptron (MLP) consists of an input layer with CombineEmbeddingDim (C.E.Dim) neurons, followed by a dropout layer with a rate of 0.3 for regularization, which helps stabilize the training process. Subsequently, the network includes hidden layers with 2048 and 1024 neurons, employing the Leaky ReLU function as the activation function to introduce non-linearity while mitigating the risk of vanishing gradients. Eventually, the output layer features #AnsToken units, along with a softmax activation to generate the predicted probability vector $\hat{y} \in \mathbb{R}^{\#\text{AnsToken}}$ for loss computation and downstream inference.

4.5. Loss Computation

After receiving the predicted probability vector $\hat{y} \in \mathbb{R}^{\#\text{AnsToken}}$ from the classification model, we compute the training loss. A dictionary is first constructed by sorting words numerically and alphabetically, tokenizing all words or phrases present in the answers from the training, validation, and test sets, as shown in Table 2. We apply one-hot encoding to the tokenized target output y_{Text} to create the one-hot target vector $\mathbf{y}$, which is required for the loss computation.

Original Word/Phrase	Token	One-hot Encoding
”11-20mm”	0	[1, 0, 0, . . . , 0]
”11-20mm, >20mm”	1	[0, 1, 0, . . . , 0]
⋮	⋮	⋮
”Z-line”	399	[0, 0, 0, . . . , 1]

Table 2. Illustration of the tokenization and one-hot encoding process for words and phrases in the dataset. Each word or phrase is assigned a unique token, subsequently converted into a one-hot encoded vector for efficient processing during model training.

The cross-entropy loss function is utilized to compute the training loss for a batch of inputs. For a batch of size N, the mean cross-entropy loss $\mathcal{L}_{\text{batch}}$ is calculated as follows:

$$\mathcal{L}_{\text{batch}} = -\frac{1}{N}\sum_{l=1}^{N}\sum_{i=1}^{\#\text{AnsToken}} y_i^{(l)} \log(\hat{y}_i^{(l)}) \tag{8}$$

where $\mathbf{y}^{(l)}$ is the one-hot encoded target vector, and $\hat{\mathbf{y}}^{(l)}$ is the predicted probability vector for the l-th input.

4.6. Textual Answer Generation

During the inference phase, the predicted probability vector $\hat{y} \in \mathbb{R}^{\#\text{AnsToken}}$ is employed to generate textual answers, utilizing either a single model approach or an ensemble method, as illustrated in Figure 3.

4.6.1 Single Model Answer Generation

In the single model approach, the predicted probability vector $\hat{y} \in \mathbb{R}^{\#\text{AnsToken}}$ is processed through the $\arg\max$ function to identify the index of the highest probability value. This index is used to select the most likely token corresponding to the expected word or phrase. Subsequently, the output answer is constructed by mapping the Token_Index to its respective word or phrase within the vocabulary dictionary. This process can be formulated as the following equations:

$$\text{Token\_Index} = \arg\max_i(\hat{y}_i) \tag{9}$$

$$\text{OutputAnswer} = \text{VocabDictionary}[\text{Token\_Index}] \tag{10}$$

where $\hat{y}_i$ represents the probability associated with the i-th token in the vocabulary.

4.6.2 Ensemble-Based Answer Generation

In the ensemble-based approach, predicted probability vectors from multiple Med-VQA systems are aggregated. Each system's output is linearly weighted, and the combined result is then passed through the $\arg\max$ function to extract the token index for the final answer as illustrated in Figure 4. This process is represented by the following equations:

$$\bar{\hat{y}} = \sum_{k=1}^{M} w_k \hat{y}_k \tag{11}$$

$$\text{Token\_Index} = \arg\max_i(\bar{\hat{y}}_i) \tag{12}$$

$$\text{OutputAnswer} = \text{VocabDictionary}[\text{Token\_Index}] \tag{13}$$

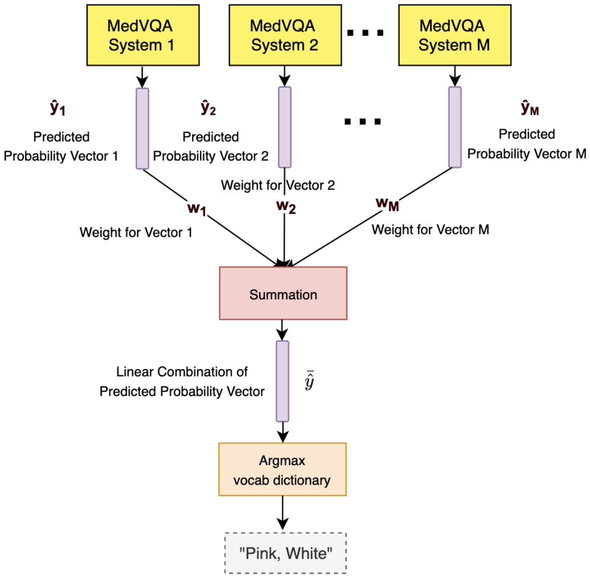


Figure 4. Architecture of the ensemble-based inference approach for answer generation. Predicted probability vectors from individual Med-VQA systems are summed and the resulting vector is used for selecting the corresponding word or phrase answer.

where $\hat{y}_k$ is the predicted probability vector from the k-th Med-VQA system, w_k denotes the weight assigned to each system, and M is the total number of systems included in the ensemble. By leveraging the combined strengths of individual models, the ensemble approach enhances prediction reliability and accuracy. In our system setup, we have chosen to assign equal weights to each system, setting $w_k = \frac{1}{M}$.

5. Experiments

In this section, we will conduct three experiments aimed at identifying the optimal model architecture with varying image feature extractors and feature fusion algorithms (Sec. 5.1), assessing the impact of regularization on model performance (Sec. 5.2), and determining the most effective ensemble model (Sec. 5.3) with the following condition, as shown in Table 3. We will train a total of eight models and construct an optimal ensemble model from these.

Control Variables and Hyperparameters
Training Set: 1600 images
Validation Set: 200 images
Test Set: 200 images
Word Encoder: BERT
Classification Model: MLP
Training Epochs: 100
Learning Rate: 1e-4
Validation Frequency: 5
Batch Size: 32
L2 Regularization: 1e-4 (For Experiment 1 only)
Loss Function: Cross Entropy
Optimizer: Adam

Table 3. Summary of the control variables and fixed hyperparameters applied in the Med-VQA experiments 1 and 2. These parameters ensure consistent model training and evaluation conditions across all tests.

5.1. Performance Comparison between Different Model Architectures

Experiment Objective : The objective of this experiment is to explore how different image feature extractors = {Resnet50, VGG16} and feature fusion algorithms = {Concatenation, Element-wise Multiplication} influence model performance, thereby identifying the most effective architecture for the Med-VQA system.

Experiment Description : In this experiment, we will construct four models, utilizing the Cartesian product of image feature extractors {ResNet50, VGG16} and feature fusion algorithms {Concatenation, Element-wise Multiplication}, as depicted from row 1 to 4 in Table 4. To ensure a consistent evaluation, a controlled experimental setup will be implemented. We will utilize the same datasets, comprising 1,600 training images, 200 validation images, and 200 testing images, as specified in Sec. 3.1. Aside from the variations in image feature extractors and feature fusion algorithms, each model will share a consistent architecture. Specifically, they will incorporate the BERT word encoder, as described in Sec. 4.2, and use a Multi-Layer Perceptron (MLP) as the classification model, as outlined in Sec. 4.4.

Dependent Variable					
Model ID	**Image Model**	**Feature Fusion Algorithm**	**Accuracy**	**BLEU-1 Score**	**Regularization**
1	ResNet50	Elementwise Multiplication	83.64%	0.8501	1e-4
2	VGG16	Elementwise Multiplication	81.67%	0.8173	1e-4
3	ResNet50	Concatenation	82.31%	0.8310	1e-4
4	VGG16	Concatenation	80.39%	0.8173	1e-4
5	ResNet50	Elementwise Multiplication	81.75%	0.8405	0
6	VGG16	Elementwise Multiplication	81.36%	0.8259	0
7	ResNet50	Concatenation	80.58%	0.8244	0
8	VGG16	Concatenation	80.14%	0.8186	0
Best Ensemble (1,4,7)	—	—	**85.92%**	**0.8635**	—

Table 4. Summary of Experiment Results: This table presents the accuracy and BLEU-1 scores for all models developed across the experiments. Rows 1 to 4 correspond to models from Experiment 1, Rows 5 to 8 to models from Experiment 2, and the final row represents the best ensemble model from Experiment 3. The row in bold text indicates the best performance.

Each model will undergo training for 100 epochs with a batch size of 32, a learning rate of 1e-4, a validation frequency of 5, and an L2 regularization strength of 1e-4. Cross-entropy loss will be employed for computing the loss, and weights will be updated via the Adam optimizer, as summarized in Table 3. Model performance will be assessed using accuracy and BLEU-1 Score.

Experiment Discussion : The analysis begins by comparing the performance of the Med-VQA systems utilizing ResNet50 with those employing VGG16. According to from row 1 to 4 in Table 4, Model 1 and Model 3, both leveraging ResNet50, respectively demonstrate accuracies of 83.64% and 82.31%, resulting in an overall average accuracy of 82.97%. Conversely, Model 2 and Model 4, both utilizing VGG16, respectively demonstrate accuracies of 81.67% and 80.39%, resulting in an overall accuracy of 81.03%. These findings suggest that the Med-VQA system based on ResNet50 outperforms its VGG16 counterpart. This disparity may be attributed to the residual connections within each block to enhance gradient flow by allowing the gradient to bypass one or more layers. This design enables the training of deeper networks without performance degradation. Consequently, it can be concluded that the Med-VQA system with ResNet50 as the image encoder potentially offers better performance over that with VGG16.

The next comparison focuses on the feature fusion algorithms, including the Element-wise Multiplication and Concatenation. Based on Table 4, Model 1 and Model 2, employing Element-wise Multiplication, achieve accuracies of 83.63% and 81.67% respectively, yielding an overall accuracy of 82.66%. In contrast, Model 3 and Model 4, utilizing the Concatenation method, attain accuracies of 82.31% and 80.39%, resulting in an average of 81.35%. These results indicate that the Med-VQA system incorporating the element-wise multiplication algorithm outperforms the systems using concatenation. This improved performance may be linked to element-wise multiplication's ability to directly merge two feature sets by multiplying corresponding elements, thereby highlighting feature interactions more effectively than simple concatenation. This operation fosters discriminative feature learning by emphasizing component-level interactions Consequently, it can be concluded that the element-wise multiplication algorithm likely enhances Med-VQA system performance over concatenation.

5.2. Assessment of Regularization Impact on Model Performance

Experiment Objective : The aim of this experiment is to evaluate the impact of weight decay on the performance and stability of the models. By replicating the setup of Experiment 1 but setting weight decay to zero, we aim to assess model performance in the absence of weight decay regularization.

Experiment Description : In this experiment, we employ the four models developed in Sec. 5.1, training them without weight decay, as detailed in rows 5 to 8 of Table 4. This setup allows us to observe the effects of eliminating weight decay on model behavior and performance metrics. To ensure consistent evaluation, we use the same control setup described in Sec. 5.1, with regularization strength as the independent variable. Model performance will be assessed using accuracy and BLEU-1 score metrics.

Experiment Discussion : We have evaluated the performance of VQA systems with and without regularization. According to Table 4, Models 1, 2, 3, and 4, which incorporate regularization, demonstrate average accuracies of 83.64%, 83.64%, 83.64%, and 83.64%, respectively, leading to an overall average accuracy of 82.00%. In contrast, Models 5, 6, 7, and 8, which do not utilize regularization, show accuracies of 81.75%, 81.36%, 80.58%, and 80.14%, resulting in an overall average accuracy of 80.96%. These results indicate that the VQA systems employing weight decay for regularization outperform those without. This improvement can be attributed to the regularization penalty, which limits excessive weight magnitudes, thereby preventing overfitting to the training data and enhancing generalization to the testing data. From these observations, it can be concluded that models utilizing regularization are likely to achieve superior performance.

5.3. Constructing the Optimal Ensemble Model

Experiment Objective : The final experiment aims to identify the optimal ensemble model by exploring all possible combinations of the eight trained models from Sections Sec. 5.1 and Sec. 5.2. Additionally, we seek to investigate whether the ensemble technique effectively synthesises predictions from multiple models, thereby enhancing overall reliability and accuracy.

Experiment Description :

8-bit Binary String	Decimal ID	Representation
00000001	1	Use Model 1
00000011	2	Use Model 1 and 2
...	...	...
11111111	255	Use All Models

Table 5. Mapping Table for different representation for indicating the composition of the ensemble model. This mapping can help us to better visual the result.

With the development of 8 trained models as detailed in Sec. 5.1 and Sec. 5.2, there exist a total of $(2^8 - 1 = 255)$ potential combinations for constructing an ensemble VQA system (assume that at least 1 model must exist in the ensemble VQA system). To manage these combinations, an 8-bit binary string is utilised to encode the ID of each ensemble configuration. In this encoding scheme, a bit value of 1 at the k-th position indicates the inclusion of model-k in the ensemble. Subsequently, the binary string can be converted into a decimal value to uniquely encode the ID of the ensemble model, as illustrated in Table 5. Finally, we can conduct a combinatorial search to test for all the possible combinations of the ensemble system with the testing set used in both Sec. 5.1 and Sec. 5.2. Model performance will be assessed using accuracy and BLEU-1 Score metrics.

Experiment Discussion :

Figure 5 indicates that the ensemble model with a combination ID of 73 (in decimal) exhibits the best performance. The corresponding 8-bit binary string for this model combination is 01001001, which signifies the inclusion of model 1, model 4, and model 7. Table 4 illustrates that the ensemble model with the optimal configuration achieves an accuracy of 85.92% and a BLEU-1 Score of 0.8635, surpassing the performance of individual VQA systems. This enhanced performance can be attributed to the ensemble's ability to mitigate bias towards the training set by averaging predictions across multiple models. This reduces the influence of outliers and prevents overfitting, thereby improving generalisation to new data. Consequently, it can be concluded that the ensemble model outperforms individual models.

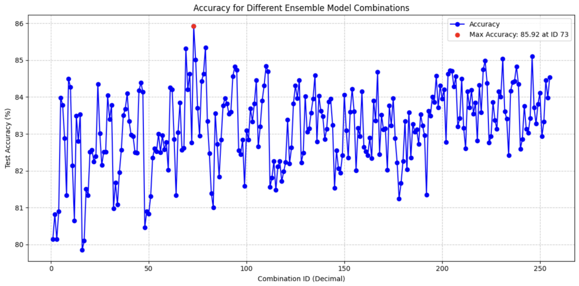


Figure 5. Plot of Model accuracy against ensemble models with different combinations.

6. Conclusion and Prospect

In this paper, we investigated the performance of a VQA model using different image encoders (ResNet50 and VGG16), feature fusion strategies (concatenation and element-wise multiplication), and regularization techniques (weight decay). After that, we evaluated the impact of an ensemble approach on overall performance. Based on the experiments presented in Sec. 5, we conclude that the Med-VQA system employing ResNet50 as the image encoder and element-wise multiplication for feature fusion achieves superior performance. Furthermore, weight decay regularization effectively mitigates overfitting, improving generalization to the test set. Moreover, the ensemble approach provides further performance gains.

Despite these advancements, there are promising directions for future research. Incorporating transformer-based models for feature fusion may significantly boost performance, albeit with increased computational demands. Transitioning to real-time systems could enable functionalities such as object segmentation and localization, which are critical for applications requiring precise detection and tracking. Exploring alternative language models for encoding textual queries, or leveraging text generation techniques rather than classification methods, could enhance the model's flexibility and accuracy in generating responses. This shift would enable handling a broader range of queries and producing more nuanced answers.

References

[1] Ali Alammari, Abm Rezbaul Islam, JungHwan Oh, Wallapak Tavanapong, Johnny Wong, and Piet C. de Groen. Classification of ulcerative colitis severity in colonoscopy videos using cnn. *Proceedings of the 9th International Conference on Information Management and Engineering*, page 139–144, Oct 2017. 1

[2] J. et al. Bertels. Ensemble learning techniques for breast cancer detection from diagnostic mammograms. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020. 2

[3] Dmitrii Bychkov, Nina Linder, Riku Turkki, Stig Nordling, Panu E. Kovanen, Clare Verrill, Margarita Walliander, Mikael Lundin, Caj Haglund, and Johan Lundin. Deep learning based tissue analysis predicts outcome in colorectal cancer, Feb 2018. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[5] S. A. Hicks, A. Storås, P. Halvorsen, T. de Lange, M. A. Riegler, and V. Thambawita. Overview of imageclef medical 2023 – medical visual question answering for gastrointestinal tract. In *CLEF2023 Working Notes, CEUR Workshop Proceedings*, Thessaloniki, Greece, 2023. CEUR-WS.org. 2

[6] S. et al. Li. Ensemble learning for image segmentation of renal tumors. In *Springer Proceedings in Advanced Robotics*, 2019. 2

[7] Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. Medical visual question answering: A survey, Jun 2023. 1

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 4

[9] Dineen Smith, Georgios I. Papachristou, and Asif Khalid. Tu1330 impact of monitoring colonoscopy quality indicators on endoscopist performance and colonoscopy quality. *Gastrointestinal Endoscopy*, 77(5), May 2013. 1

[10] Constantine P. Spanos. Malignant polyps. *Colorectal Disorders and Diseases*, page 233–236, 2023. 1, 2

[11] T. M. Thai, A. T. Vo, H. K. Tieu, L. N. Bui, and T. T. Nguyen. Uit-saviors at medvqa-gi 2023: Improving multimodal learning with image enhancement for gastrointestinal visual question answering. In *CLEF2023 Working Notes, CEUR Workshop Proceedings*, Thessaloniki, Greece, 2023. CEUR-WS.org. 2

[12] World Health Organization. Colorectal cancer fact sheet, 2020. Accessed: 2024-11-11. 1, 2